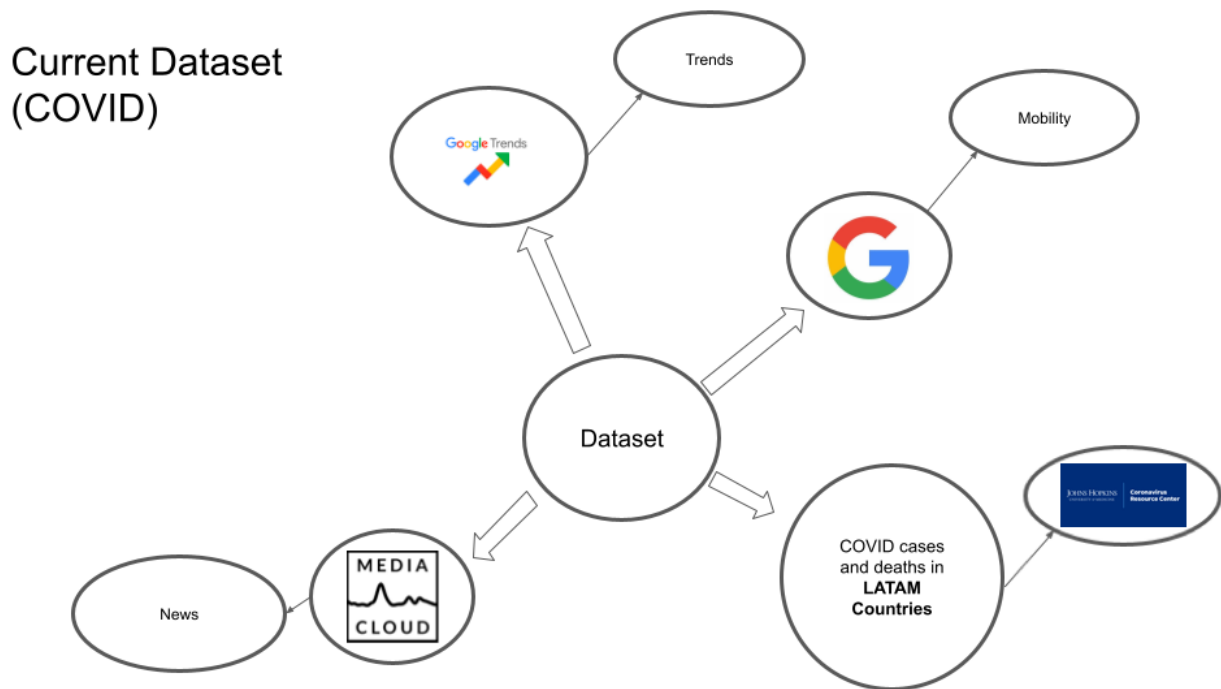# Infodemic and misinformation in the spread of COVID-19 in LATAM.

The dataset is composed by 3 different sources.
1. COVID-19 data in LATAM from Our World in Data
2. Internet sources which are:
      2.1. Media Cloud
      2.2. Google Trends

NOTE: For more information on the dataset and codes used during its generation, consult the repository: https://github.com/MITCriticalData-Colombia/Covid_Media_Cloud

1. COVID-19 data in LATAM
The LATAM COVID data is made up of 7 different datasets:

- **covid_latam(raw).csv**: Contains COVID data for 32 LATAM countries. The dataset contains the columns:
    - 'iso_code': ISO country code in 3 characters.
    - 'continent': Continent where the data comes from.
    - 'location': Country where the data comes from.
    - 'date': Date on which the data was recorded.
    - 'total_cases': Total covid cases (cumulative) or probable cases.

- 'new_cases': New covid cases presented in the country and established date.
- 'New_cases_smoothed': New confirmed cases approximating the current week.
- 'Total_deaths': Total deaths attributed to covid-19.
- 'New_deaths': new deaths attributed to covid-19
- 'New_deaths_smoothed': new covid-19 deaths approximating the current week.
- 'Total_cases_per_million': Total confirmed cases of COVID-19 per 1,000,000 people.
- 'New_cases_per_million': New confirmed cases of COVID-19 per 1,000,000 people.
- 'New_cases_smoothed_per_million': New confirmed cases of COVID-19 per 1,000,000 people approximating the current week.
- 'Total_deaths_per_million': Total deaths attributed to COVID-19 per 1,000,000 people
- 'New_deaths_per_million': New deaths attributed to COVID-19 per 1,000,000 people.
- 'New_deaths_smoothed_per_million': New deaths attributed to COVID-19 approximating the current week per 1,000,000 people.
- 'Reproduction_rate': Real-time estimate of the effective reproduction rate (R) of COVID-19, that is, this measure is how fast the virus is spreading in a population at a given time.
- 'Icu_patients': Number of patients with COVID-19 in intensive care units (ICU) on a given day.
- 'Icu_patients_per_million': Number of COVID-19 patients in intensive care units (ICU) on a given day per 1,000,000 people
- 'Hosp_patients': Number of COVID-19 patients in the hospital on a given day.
- 'hosp_patients_per_million': Number of COVID-19 patients in the hospital on a given day per 1,000,000 people.
- 'Weekly_icu_admissions': Number of COVID-19 patients recently admitted to intensive care units (ICUs) in a given week.
- 'Weekly_icu_admissions_per_million': Number of COVID-19 patients recently admitted to intensive care units (ICUs) in a given week per 1,000,000 people.
- 'Weekly_hosp_admissions': Number of newly admitted COVID-19 patients in hospitals in a given week.
- 'weekly_hosp_admissions_per_million': Number of newly admitted COVID-19 patients in hospitals in a given week per 1,000,000 people
- 'total_tests': Total tests for COVID-19
- 'New_tests': New tests for COVID-19 (only calculated for consecutive days).
- 'Total_tests_per_thousand': Total tests for COVID-19 per 1,000 people.
- 'New_tests_per_thousand': New tests for COVID-19 per 1,000 people.
- 'new_tests_smoothed': New tests for COVID-19 approximating the current week. For countries that do not report testing data on a daily basis, we assume that testing changed equally on a daily basis and are averaged over a 7-day rolling window.
- 'New_tests_smoothed_per_thousand': New tests for COVID-19 approximating the current week per 1,000 people
- 'positive_rate': The proportion of COVID-19 tests that are positive, given as a 7-day moving average (this is the inverse of tests_per_case).

- 'tests_per_case': Tests performed per new confirmed case of COVID-19, given as a 7-day rolling average (this is the inverse positive_rate)
- 'tests_units': Units used by the location to report its test data.
- 'Total_vaccinations': Total number of COVID-19 vaccination doses administered.
- 'people_vaccinated': Total number of people who received at least one dose of vaccine.
- 'People_fully_vaccinated': Total number of people who received all the doses prescribed by the initial vaccination protocol.
- 'Total_boosters': Total number of COVID-19 vaccination booster doses administered.
- 'new_vaccinations': New COVID-19 vaccination doses administered (only calculated for consecutive days).
- 'New_vaccinations_smoothed': New COVID-19 vaccination doses administered, given as a 7-day rolling average.
- 'total_vaccinations_per_hundred': Total number of COVID-19 vaccination doses administered per 100 people in the total population.
- 'People_vaccinated_per_hundred': Total number of people who received at least one dose of vaccine per 100 people in the total population.
- 'people_fully_vaccinated_per_hundred': Total number of people who received all the doses prescribed by the initial vaccination protocol per 100 people in the total population.
- 'Total_boosters_per_hundred': Total number of booster doses of vaccination against COVID-19 administered per 100 people in the total population.
- 'New_vaccinations_smoothed_per_million': New doses of COVID-19 vaccination administered (7 days smoothed) per 1,000,000 people in the total population.
- 'New_people_vaccinated_smoothed': Daily number of people receiving their first dose of vaccine (7 days smoothed).
- 'new_people_vaccinated_smoothed_per_hundred': Daily number of people receiving their first dose of vaccine (smoothed 7 days) per 100 people in the total population.
- 'Stringency_index': Government Response Stringency Index: Composite measure based on 9 response indicators including school closures, workplace closures and travel bans, rescaled to a value from 0 to 100 (100 = response plus strict)
- 'Population': Population (last available values).
- 'population_density': Number of people divided by land area, measured in square kilometers, most recent year available
- 'median_age': Median age of the population, UN projection for 2020
- 'Aged_65_older': Proportion of the population aged 65 or older, most recent year available
- 'aged_70_older': Percentage of the population aged 70 or older in 2015
- 'gdp_per_capita': Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available.
- 'Extreme_poverty': Proportion of the population living in extreme poverty, most recent year available since 2010.
- 'cardiovasc_death_rate': Mortality rate due to cardiovascular disease in 2017 (annual number of deaths per 100,000 people).

- 'diabetes_prevalence': Prevalence of diabetes (% of the population aged 20 to 79 years) in 2017.
- 'Female_smokers': Proportion of women who smoke, most recent year available.
- 'male_smokers': Percentage of men who smoke, most recent year available.
- 'Handwashing_facilities': Proportion of population with basic handwashing facilities in facilities, most recent year available.
- 'Hospital_beds_per_thousand': Hospital beds per 1,000 people, most recent year available since 2010.
- 'Life_expectancy': Life expectancy at birth in 2019.
- 'Human_development_index': A composite index that measures average achievement in three basic dimensions of human development: a long and healthy life, knowledge, and a decent standard of living.
- 'excess_mortality_cumulative_absolute': Cumulative difference between the reported number of deaths since January 1, 2020, and the projected number of deaths for the same period based on previous years.
- 'Excess_mortality_cumulative': Percentage difference between the cumulative number of deaths since January 1, 2020 and the projected cumulative deaths for the same period based on previous years.
- 'excess_mortality': Percentage difference between the reported number of weekly or monthly deaths in 2020-2021 and the projected number of deaths for the same period based on previous years.
- 'Excess_mortality_cumulative_per_million': Cumulative difference between the reported number of deaths since January 1, 2020, and the projected number of deaths for the same period, based on previous years, per million people.

Note: For more details regarding the calculation of the excess mortality for the different countries, please refer to: https://elifesciences.org/articles/69336

- **Daily_covid_cases_latam.csv**: Daily COVID cases per million inhabitants for 32 LATAM countries.
  - The first 'location' column contains the name of the country where the cases were filed.
  - The following columns contain the number of cases per million inhabitants presented on that day in that country.

- **daily_covid_deaths_latam.csv**: Daily confirmed COVID deaths per million inhabitants for 32 LATAM countries.
  - The first column 'location' contains the name of the country where the deaths occurred.
  - The following columns contain the number of confirmed deaths per million inhabitants presented on that day in that country.

- **daily_covid_vaccinations_latam.csv**: Daily confirmed COVID vaccinations per hundred inhabitants for 32 LATAM countries (in a cumulative way).
  - The first column 'location' contains the name of the country where the vaccinations were applied.

- The following columns contain the number of confirmed vaccinations per hundred inhabitants applied till that day in that country.

- **Weekly_covid_cases_latam.csv**: Weekly COVID cases using the epidemiological week per million inhabitants for 32 LATAM countries.
  - The first 'location' column contains the name of the country where the cases were filed.
  - The following columns contain the sum of cases per million inhabitants presented in that country during that epidemiological week. The format of the epidemiological week is:
    - First 4 digits corresponding to the year, followed by 2 digits corresponding to the epidemiological week.

- **Weekly_covid_deaths_latam.csv**: Weekly confirmed COVID deaths using the epidemiological week per million inhabitants for 32 LATAM countries.
  - The first column 'location' contains the name of the country where the deaths occurred.
  - The following columns contain the sum of the deaths per million inhabitants presented in that country during that epidemiological week. The format of the epidemiological week is:
    - First 4 digits corresponding to the year, followed by 2 digits corresponding to the epidemiological week.
- **weekly_covid_vaccinations_latam.csv**: Weekly confirmed COVID vaccinations applied using the epidemiological week per hundred inhabitants for 32 LATAM countries.
  - The first column 'location' contains the name of the country where the vaccinations were applied.
  - The following columns contain the sum of the total vaccinations per hundred inhabitants presented in that country till that epidemiological week. The format of the epidemiological week is:
    - First 4 digits corresponding to the year, followed by 2 digits corresponding to the epidemiological week.

Note: For weekly documents, each column represents que epiweek in epiweek format (Year-Week). You can use the epiweek library of python to convert the columns to date format.
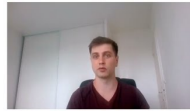
2. Internet sources:

2.1. Media Cloud

Media cloud dataset was extracted using Explorer tool from Media Cloud. Some examples of the use of the tool can be seen in the videos:

1. [Media Cloud - Topic Mapper](#):

2. [Media Cloud - Explorer](#):

The datasets that we find in this section are:
- **covid-19_attention_latam.csv**: Attention to the word COVID-19 in LATAM countries.
  - The first column 'date' contains the date on which attention to the word is being analyzed.
  - For each country, we have the following features of the word "covid-19" in each country:
    - "count": which represents the number of news media mentioning our topic.
    - "total_count" which is the total number of news media published.

- ■ and "ratio" which indicates the proportion of news media that contain the word we are interested in.
  - ○ The column is presented in the format:
    - ■ count_NameOfCountry, total_count_NameOfCountry, ratio_NameOf-Country

- **covid_vaccination_attention_latam.csv**: Attention to the word COVID vaccination in LATAM countries.
  - ○ The first column 'date' contains the date on which attention to the word is being analyzed.
  - ○ For each country, we have the following features of the word "COVID vaccination" in each country:
    - ■ "count": which represents the number of news media mentioning our topic.
    - ■ "total_count" which is the total number of news media published.
    - ■ and "ratio" which indicates the proportion of news media that contain the word we are interested in.
  - ○ The column is presented in the format:
    - ■ count_NameOfCountry, total_count_NameOfCountry, ratio_NameOf-Country

2.2. Google Trends

The Google Trends dataset was extracted using the Google Trends Online tool. Some examples of the use of the tool can be seen in the video:

2. [Google Trends tutorial](#):

The dataset that we find in this section is:

- **covid-19_Symptoms_latam.csv**: Google trends of some Covid-19 symptoms in LATAM countries.
    - The first column 'Week' contains the date on which symptom trends in LATAM are being analyzed.
    - Each country has Fever, Headache and Cough columns, which represent some of the symptoms of covid-19 in each country. The column is presented in the format:
        - Fever_NameOfCountry, Headache_NameOfCountry, Cough_NameOf-Country